THE STATISTICAL METHOD IN PROBLEMS OF WATER SUPPLY QUALITY

By Abel Wolman, *Maryland State Department of Health*

INTRODUCTORY

The concept of water supply quality has the simplicity of the unknown to the layman, but the complexity of the universe to the sanitarian. If one uses the mathematician's measure of the complexity of a function—the number of its attributes—the problem of water supply quality, a function dependent upon mutually active natural, physical, chemical and biological phenomena, offers an attractive field of study to the statistician. For the professional statistician has been concerned always with "quantitative data affected by a multiplicity of causes"[1] and with their elucidation. In considering the causes operating to produce relatively good or bad waters, such as rainfalls, pollution, purification, etc., and their interpretation upon the basis of laboratory findings and personal surveys it becomes manifest that problems of water supply quality fall well within the scope of statistical method. Just as in all statistical problems, so in that of water supply quality, the investigator is confronted with the two-fold task of determining the method of evaluating the units of interpretation and of defining the limiting values of such units. The method of approach to each problem involves a statistical viewpoint, as well as a quantitative methodology. The present paper has been prepared in order to illustrate, in as brief terms as possible, this statistical method of approach, by developing therein a few examples of its application to the question of water supply quality. The writer plans to trace the evolution of the concept of water supply quality in the sanitarian's mind and to point out in such a development the function which the statistical art has performed or may be expected to supply in the future. The discussion appears to be a necessary one since hitherto the water supply investigator has been accused of an aversion for the quantitative sciences, while, on the other hand, the professional statistician has shown a neglect of a field which perhaps did not appear to be worthy of his mettle. The present study may serve to remove this friendly distrust which retards in a degree progress in critical studies of water supply quality.

I. THE LABORATORY EXAMINATION OF WATER SUPPLIES

The sanitary quality of water supply must be predicated necessarily upon the demonstration of its relative inability to produce disease.

With the present germ theory of disease, such a demonstration resolves itself into the laboratory problem of enumerating the number and types of pathogenic organisms in stated quantities of water. It is apparent, therefore, that the technique in this instance is largely bacteriological, and the discussion, for purposes of simplicity, may be restricted to the problems of the evaluation of bacterial units, as illustrative of the statistical method.

It is manifestly impossible and impracticable to examine an unknown water in such manner as to determine its content of all kinds of bacteria or even of those relatively few classes of specific organisms which it is known are both disease-producing and capable of living in water. It is more desirable as well as convenient, therefore, to choose one family or group of micro-organisms whose natural habitat and life history are similar to the variety of pathogenic organisms and whose detection by laboratory methods is most simple and speedy. Bacteriologists have concluded that a particular class of bacteria serves as the most convenient index to water supply quality or contamination. They have chosen as this index class or type the so-called colon or bacillus coli group. The B. coli group has been so selected, because its origin is in general the colon or digestive tract of man and its presence is usually indicative of human sewage pollution (the possible and probable existence of colon types in other environments need not concern us at this point).

One of the primary objects of the bacteriologist, therefore, is to differentiate the bacterial species present in a water supply, so as to demonstrate the presence or absence of members of the B. coli group. In addition, it is necessary to obtain some idea of the relative frequency of such a group, since smaller numbers naturally connote a more remote pollution, due to the dying off of bacteria in the unfavorable environment of water, to the presence of antagonistic life, and to other natural and artificial barriers to its development. The problems arising in the laboratory differentiation of bacterial varieties offers, therefore, material for an initial example. Two general methods of distinguishing groups of bacteria are available. Both are based upon the method of differences. In the one case, morphological or structural characteristics, and in the other, metabolic distinctions control. Various classifications of the colon group, for instance, are based upon its ability to produce acid and gas from fermentable substances. Investigators have observed that certain types of B. coli ferment such complex organic compounds as sucrose, dulcitol and raffinose while others do not. Differences in the amount and character of gas formation from certain substances distinguish other types of bacteria.

In all classifications, however, it has been recognized that the same group may have a variety of reactions which overlap partially those of other groups. Two types of bacteria, for instance, may both ferment sucrose, but may differ in their effect upon a second or third compound. This gives rise naturally to a vast amount of possible combinations between characters and Levine[2] points out that "as the number of fermentable substances increases, the number of varieties increases geometrically approaching infinity. The number of 'varieties' is given by the formula 2^n where 'n' is the number of characters studied. Thus with 8 characters there are 256 possible combinations; this number rises to 1,024 with 10 characters and to 65,536 when 16 characters are observed. The absurdity of regarding each character as of similar and equal differential value is thus evident."

Levine, as well as other more recent investigators, has concluded that the principle of the correlation of characters should be emphasized in the attempt to distinguish bacterial species. He points out that certain properties have been universally accepted, after long checking, as reliable evidences of bacterial differences. Among such properties, he enumerates the selective dyeing of bacteria, their powers of spore formation, and their adaptation to aerobic or anaerobic development. The taxonomic value of the characters of motility, indol formation, and fermentation of certain compounds, on the other hand, he assumes to be still debatable. In order to avoid the adoption of a confusing classification of bacteria upon the basis of every character studied (of which we have indicated only a few) he has recourse to a basis of subdivision "on that character which gives the greatest amount of information as to the manner in which the resulting sub-groups react with respect to other characters."[2] By making use of the above principle Levine evolves a classification of coli-like bacteria which is based almost completely upon statistically evaluated correlated characters. For the purpose of this study, he recognizes two main strains of bacteria, the B. coli and the B. aerogenes-cloacae group, which earlier investigations have shown to be distinguishable most often by their reactions to methyl-red and to the Voges-Proskauer reagent. The first strain is usually methyl-red positive and Voges-Proskauer negative, while the second strain shows the reverse. The justification of this initial subdivision into two main groups consists in the fact that the strains thus subdivided show end products of carbohydrate fermentations of two entirely distinct kinds.

Levine's procedure consists in tabulating all of the reactions of the organisms studied in each of the above two groups in two different tables, from which are calculated the coefficients of correlation for each

pair of characters. He selects, then, for subdivision that character which gives the highest coefficient of correlation with the greatest number of other characters. For these resulting sub-groups new correlation tables are prepared and further subdivision is made. These sub-groups are regarded as species and each is assigned its name.

In order to illustrate Levine's use of the coefficient of correlation for taxonomic purposes, let us follow his procedure in the subdivision of the B. coli, or methy-red positive and Voges-Proskauer negative, group of bacteria. For the 182 strains of this group that were studied by means of microscopic and metabolic methods, the coefficients of correlation shown in Table I were obtained.

TABLE I.

COEFFICIENTS OF CORRELATION OBTAINED FROM PAIRS OF CHARACTERS AMONG 182 STRAINS OF THE B. COLI GROUP

	Motility	*Indol*	*Sucrose*	*Raffinose*	*Dulcitol*	*Glycerol*	*Salicin*
Motility		−.39	+.53	+.43	+.53	+.18	+.40
Indol	−.39		+.08	+.00	+.02	−.28	+.76
Sucrose	+.53	+.08		+.99	+.58	−.38	+.20
Raffinose	+.43	+.00	+.99		+.58	−.29	+.27
Dulcitol	+.53	+.02	+.58	+.58		−.21	+.60
Glycerol	+.18	−.28	−.38	−.29	−.21		+.52
Salicin	+.40	+.76	+.20	+.27	+.60	+.52	

Since Levine's criterion for the choice of a character for subdivision is that that character should give the highest coefficient of correlation with most other characters, it is apparent, from an inspection of Table I, that sucrose, raffinose, dulcitol, and salicin meet this criterion more completely than do other properties. For special technical reasons, Levine chooses sucrose for primary division of the B. coli group and obtains by differentiation on sucrose ninety-three strains of the sucrose positive and eighty-nine strains of the sucrose negative groups. These two groups combined form, of course, the total of 182 strains initially chosen for study. Further study of the sucrose positive strains discloses a series of coefficients of correlation of characters as shown in Table II.

TABLE II.

COEFFICIENTS OF CORRELATION FOR PAIRS OF CHARACTERS AMONG 93 SUCROSE POSITIVE STRAINS OF THE B. COLI GROUP

	Motility	*Indol*	*Dulcitol*	*Glycerol*	*Salicin*
Motility		−.27	+.67	+.40	+.54
Indol	−.27		+.05	−.42	+.28
Dulcitol	+.67	+.05		−.32	+.39
Glycerol	+.40	−.42	−.32		+.32
Salicin	+.54	+.28	+.39	+.32	

Table II indicates that motility is the best correlated character and this property provides, therefore, for two further sub-groups, a sucrose-positive motile sub-group and a sucrose-positive non-motile group. These sub-groups are treated in the manner already illustrated and the coefficient of correlation for different characters provide for further subdivision. With the aid of this statistical interpretation of his studies of 333 coli-like bacteria, isolated from various sources, Levine suggests a classification of bacterial varieties. The summary of this classification need not be repeated here, since the reader is interested more in his method of attack than in the resulting bacteriological findings.

Such classifications as Levine's supply the sanitarian with the qualitative information necessary for the interpretation of one phase of the water supply quality problem.* The analyst dealing with waters is concerned not only with the nature of the bacterial types present therein, but also in the magnitude of their content, since it is the latter which indicates the degree and the remoteness of pollution. In the search for a potable water, it is often useless to seek that water which has no possible source of contamination, but it is always necessary to determine the quantitative bacterial importance of the latter. The methods so far described answer only one question, that is, what types of bacteria are present in the water. In the solution of the second inquiry, regarding the number of a particular type in a stated quantity of water, statistical method has played recently an important part.

In the simpler tests for the B. coli group in waters, the so-called fermentation-tubes are used. These tubes contain the medium selected for most efficient differentiation of the B. coli group from other kinds of bacteria and are inoculated with specific quantities of the water to be tested. The production of gas in the tubes after stated periods of incubation indicates the presence of the B. coli group. Our knowledge that of five tubes, each inoculated with 0.1 c.c. of the water, four show the presence of the organism, is of value, but more important is the additional fact that such a series of findings indicates that the probable number of organisms in the sample tested is about 1,600 per 100 c.c. This conversion of qualitative fermentation-tube results into quantitative values is of special interest to the statistician.

In 1915, McCrady[3] showed that "the frequency of the appearance of the fermenting organism in the volume drawn from the sample for the test is an exponential function of the number of such organisms in the sample," and that "every fermentation-tube result, whether simple or compound, corresponds to one most probable number of organisms."

*The subdivisions Levine develops have their importance to the investigator in the fact that species or varieties appear to be somewhat correlated with habitat or source of pollution.

By employing the theory of probabilities, he demonstrates that, given the result "$\frac{p}{p+q}$ in 1 volume," for instance, the corresponding most probable number is given by the solution for x of the equation

$$1-\left(\frac{V-1}{V}\right)^x=\frac{p}{p+q}.$$

Thus, for the result "five out of ten tubes positive in 1 c.c.," the most probable number is given by solution of the equation $1-.99^x=5/10$, since $V=100$ c.c., assumed as the original quantity of water sampled. The equation being solved, $x=69$ or the most probable number of B. coli in the sample, per 100 c.c.

For compound results, such as $\frac{p}{p+q}$ in 10 c.c., $\frac{r}{r+s}$ in 1 c.c., a more complicated formula is employed which is built up, as follows:[3]

For the result $\frac{p}{p+q}$ in 10 c.c. the equation becomes $(p+q)$ (log .9) $=\frac{p\ (\log\ .9)}{1-.9^x}$ which is obtained by differentiating for a maximum the equation given in the earlier paragraph for the probability of the results.

If the result is $\frac{p}{p+q}$ in 10 c.c., $\frac{r}{r+s}$ in 1 c.c., the equation stands

$$(p+q)\ (\log\ .9)+(r+s)(\log\ .99)=\frac{p\ (\log\ .99)}{1-.9^x}+\frac{r\ (\log\ .99)}{1-.99^x}$$

where $(p+q)$ = number of tubes inoculated with 10 c.c. of sample
$(r+s)$ = number of tubes inoculated with 1 c.c. of sample
x = number of fermenting organisms in 100 c.c. of sample
p and r = number of tubes giving positive results in 10 and 1 c.c. respectively.

If lower additional quantities of water are tested, extra similar terms are added to each side of the above equation. This equation has been modified by Wolman and Weaver[4] into

$$100\ (p+q)+10(r+s)=\frac{100p}{1-.9^x}+\frac{10r}{1-.99^x}$$

since, approximately, log .9 = 10 log .99 = 100 log .999.

McCrady published later[5] a series of tables for the rapid interpretation of these results which makes the standardized use of the probable numbers of B. coli possible for the water supply investigator.*

*The assumption of McCrady that the distribution of B. coli is similar to that in a mixture of a few red balls with many white balls is to be contrasted with the hypothesis of other workers that bacteria are uniformly distributed in water (G. C. Whipple[13]). More recent independent investigators, however, confirm McCrady's assumptions.

The work of McCrady was followed by other investigations dealing with the numerical interpretations of B. coli tests, of which the more important are Stein[6], [7], [8], Greenwood and Yule[9], and Wells[10], [11], [12] The results of Stein and Greenwood and Yule, although differing in technique and in additional interesting viewpoints, are in substantial agreement with those obtained by McCrady. Stein[8] adds considerable interesting statistical material to the B. coli problem by introducing the so-called B. coli factor method, in which he considers the most probable number of B. coli per c.c. from the percentage of positive tests, the expected error of results, the study of the distribution of coli during a series of tests, and the "coli characteristic" which attempts to show by one figure, the average coli, the expected error and the variable distribution.

The discussion of the problem by Greenwood and Yule[9] has all the intricacy and mathematical complexity usually associated with Yule's contributions. Their findings, however, agree with those of McCrady and Stein. Greenwood and Yule, for instance, give as their formula for the number of B. coli per c.c., when using several tubes with 1 c.c. each

$$x = \text{B. coli per c.c.} = 2.3 \log \frac{p+q}{q}$$

whereas McCrady gives for the same condition (using an original size sample of 1,000 c.c.)

$$x = \frac{\log \frac{q}{p+q}}{1{,}000 \log .999} = \frac{\log \frac{q}{p+q}}{-1{,}000(.0004344)} = \frac{\log \frac{q}{p+q}}{-.4344} = -2.3 \log \frac{q}{p+q}$$
$$= 2.3 \log \frac{p+q}{q}$$

Perhaps the mathematician's interest may be aroused to the sanitarian's problems of water supply by the mere examination of Greenwood and Yule's discussions, while the bacteriologist may view with some alarm the same paper. It should be postulated in either case, however, that superficial considerations should not prevent the mutual aid which these two branches of science may extend to each other. While such complexity of treatment of the numerical interpretation of fermentation-tube tests as is indicated by the formula

$$P = \frac{\int_o^k \left[e^{-ha_1 n_1} \left(1 - e^{-ha_1}\right)^{m_1} \cdot e^{-ha_2 n_2} \left(1 - e^{-ha_2}\right)^{m_2} \right.}{\int_o^w \left[e^{-ha_1 n_1} \left(1 - e^{-ha_1}\right)^{m_1} \cdot e^{-ha_2 n_2} \left(1 - e^{-ha_2}\right)^{m_2} \right.}$$
$$\left. \cdots e^{-ha_n n_n} \left(1 - e^{-ha_n}\right)^{m_n} \right] dh$$
$$\left. \cdots e^{-ha_n n_n} \left(1 - e^{-ha_n}\right)^{m_n} \right] dh$$

may attract the statistician, it is hoped that it may not at the same time deter the laboratory technician from the adoption of devices which provide for more adequate solutions of his problems. Emphasis must be placed upon the fact that the mental attitude resulting from the adoption of statistical method has much promise in a field of endeavor where laboratory findings are too infrequently tested for accuracy of interpretation and rarely treated as examples of mass phenomena. The work of such men as Stein and McCrady has done much to introduce such methods by clarifying our concepts of fermentation tube results and their relative significance.

That the statistical method is an important asset in the exposition of laboratory findings is illustrated in another series of studies of various phases of water supply. Whipple[13], for instance, has demonstrated that "if, in a series of daily observations of the number of bacteria in a filter effluent extending over a year the deviation of any determination from the mean should be found to be more than five times as much as the probable error, to use a round number, this should be rejected from the series as being, for some reason or other, abnormal." He has made important contributions to the study of the frequency distributions of measures of various bacteriological, biological, and chemical characteristics of water, such as the preliminary finding that extended series of filter effluent results follow definite statistical laws in their distribution. His conclusion has been further substantiated by the more recent study of Wolman[14] of thousands of laboratory findings, in which it is indicated that the logarithms of bacterial counts, through long periods of time, have the characteristic normal probability distribution of more familiar biological statistical data.

It is of considerable interest to refer at this point to a form of graph presentation of data developed by a sanitary engineer which may be unfamiliar to most statisticians. Allen Hazen[15] in 1914 devised a form of chart ruled with a horizontal scale so divided tha the curve of probability would plot thereon as a straight line. Any series of observations, therefore, which varied in accordance with the probability law would plot also as a straight line. Illustrations of the use of such paper in water supply problems may be found in the original paper of Hazen[15] and in subsequent discussions by Whipple[13] and Wolman[14].

Stein[16], in his study of the bacterial count in water and sewage, has added considerable material to our conceptions of the variability of laboratory findings and their importance in practical studies. He has concluded, after an interesting detailed analysis of the problem, that:

(a) For platings of a single sample of water, the mean error is equal to the square root of the number of colonies on a single plate, or the square root of the average number of colonies on several plates.

(b) The variations to be expected for careful and accurate work with bacterial counts are indicated by:

(1) Standard Deviation of ± 12%

(2) Deviation (1 in 10 times) of ± 25%

For ordinary routine work:

(1) Standard Deviation of ± 25%

(2) Deviation (1 in 10 times) of ± 50%

His comparison of the characteristics of bacteriological data with certain mathematical series should be of interest to the reader, since he shows, for example, that for daily tests of Lake Erie water for one month the Lexian Ratio is 29.00 and the Disturbancy Coefficient 124.00, while the corresponding values for a normal mathematical series (Bernouilli) are given as 1.00 and 0.00 respectively.

II. THE INTERPRETATION OF THE QUALITY OF WATER SUPPLIES

In preceding paragraphs the writer has indicated a few of the problems encountered in the laboratory technique of water supply examination, which lend themselves to statistical treatment. It has been impossible to include in the present brief paper any complete survey of such applications to other phases of laboratory procedure, but sufficient material has been presented, to demonstrate that the data in the field of laboratory technique have considerable to offer to the professional statistician as bases for the development of interpretative principles of quality.

The writer believes that some mention should be made briefly of certain interesting possibilities of development in the application of statistical method to general problems of laboratory procedure. The use of the coefficient of partial correlation, for instance, does not appear to have been introduced widely in the interpretation of laboratory findings, yet the necessity for its application is most apparent. Often investigative work in water supplies is carried out on a large or plant scale with the aid of analytical laboratory methods. In the study of the chlorination of a water supply, for example, a number of different variable quantities such as turbidity, color, organic content, and bacterial densities have their effect in modifying the efficiency of the disinfection process. In practically all conclusions from such studies no attempt is made to determine mathematically the effect of such variables, other than by mere inspection of tabulated data. There is little doubt that erroneous conclusions are often obtained through the failure to evaluate quantitatively the importance of fluctuations in the various characteristics of waters subject to chlorination. It is almost impossible to determine by qualitative inspection of a series of daily observations, over an entire year, of temperature, turbidity, color, organic

content, and bacterial density in a water supply, whether the effect of a constant dosage of chlorine is influenced more greatly by any one of the above characteristics or by a combination of several or all of them.

The same problem arises, of course, in the study of any of the phenomena associated with the purification of water supplies. In the coagulation of suspended matter in water, for instance, all the variables such as time, agitation, temperature, hydrogen-ion concentration, nature of suspended matter, and character of coagulant play an interconnected part. The principle of partial correlation could be adapted with profit to these problems of associated phenomena.

The application of such a statistical principle as pointed out above is complicated, however, by the fact that the more simple statistical coefficients usually cannot be directly applied to the problems encountered, on account of the fact that such measures presuppose the use of data having a symmetrical or Gaussian distribution, while the phenomena with which the sanitarian has to deal often are characterized by asymmetrical distributions.[17], [18]

Michael[18] has discussed in this connection the determination of the most probable number of bacteria present in a sample and has demonstrated that it is not permissible to apply the probable error in the usual manner on account of the fact that the logarithms of the plate counts, and not the counts themselves, show a Gaussian frequency distribution[19]. McEwen and Michael[17] in another field of investigation have been confronted with the same problem of determining the "functional relation of one variable to each of a number of correlated variables" where such variables do not show the usual symmetrical frequency distribution. It is manifestly impossible to extend in this paper the elucidation of these applications of statistical method to problems of laboratory and plant, but the reader may find profitable data in the original papers already noted.

The opportunity for the application of statistical tests to problems of water supply quality is not restricted, however, to the materials of the analyst. The consideration of the potability of a supply involves always a series of mutually active attributes, each of which has its importance in determining the character of the water. The concept of quality connotes, therefore, a composite of properly weighted individual and fundamental units, in the evaluation of which statistics again comes to the fore.

It is unfortunate, however, that in the field of interpretation of quality statistical method has been even slower of application than in the corresponding study of laboratory data. The quantitative evaluation of sanitary data has always given way to the liberal exercise of

expert personal judgment. Where a multiplicity of causes predetermines a phenomenon, such as quality, it was thought that a proper perspective was possible only through the development of a maturity of judgment in which the play of the manifold effects was qualitatively summarized rather than quantitatively analyzed. As the methods of diagnosis of quality developed, however, the opportunity for the fruitful application of the principles of mass phenomena gradually becomes apparent. With this development of a new viewpoint, good as well as evil sometimes resulted. A complete swinging of the pendulum to the quantitative side of interpretation was feared, where the attempt was made to substitute for individual experience and judgment pseudo-quantitative measures of doubtful significance. Some of these efforts, in which statistical laws frequently were ignored, will be discussed later in this paper. In general, however, a realization is gradually coming over the sanitarian that statistics as a means, rather than as an end, has much to offer in the clarification of his problems. If the succeeding pages seem somewhat bare, in their statistical implication, the professional statistician should remember that the concepts there discussed mark the advance of a new light in sanitary engineering, which, though feeble in its flicker, gives promise of a greater brilliance in the not distant future.

Attempts to formulate water supply standards of composite character represented one of the earliest applications of semi-statistical method. Most of these were based upon the erroneous conclusion that methods of evaluating units had been standardized throughout the country. Attention has been called to this fallacy of endeavoring to establish limiting values of units attained by varying methods by Hinman[20], Norton[21], and Morse and Wolman[22]. Fundamental training in statistical interpretation no doubt would prevent the adoption of water supply quality standards before the principles of unit evaluation have been rigidly enforced.

It is not amiss, perhaps, to call attention at this point to the close analogy between the so-called scoring of a water supply, or the quantitative allocation of the quality upon the scale of sanitary safety, and the statistician's concept of index numbers. Wolman[14] has shown recently that the operations involved in making a price index number are similar to those followed, to a greater or less extent, by investigators of water supply scores. In the case of price index numbers, the object of weighing is to give each commodity included in the index number an influence upon the results corresponding to its commercial importance. In water supply index numbers, the object of weighing likewise is to give each factor making up the score an influence upon

the results corresponding to its sanitary importance. Although the problems in the two fields are the same, their solutions are necessarily different, since, in the case of water supply scores, the conversion to a common base of such units as bacterial results, sanitary surveys, operating efficiencies, etc., cannot be carried out because of the presence of varying personal opinion or judgment. It has been noted[14], however, that it still remains possible to make use nationally of simplified index numbers of water supply quality restricted in their range of significance and composed of similar units or, better still, of individual units, provided the method of evaluation of such units has been definitely and completely fixed.

Interpretations of the quality of a water include frequently more than a summary of the structural and environmental features of the supply. The possibilities of the intelligent and fruitful application of statistical devices, such as the coefficients of correlation and of variation, to other phases of water supply are mentioned only briefly here, since their complete discussion would involve a paper of a far too great length. Whipple* for instance, has suggested the use of the coefficient of correlation in analyzing the vital statistics of cities which have made changes from poor to good quality water supplies, in order to demonstrate quantitatively the existence of the Mills-Reincke phenomenon. Hazen[15] has made excellent use of statistical method in his analysis of the storage provided in an impounding reservoir on any stream and the quantity of water which can be supplied continuously by it. He introduces the coefficient of variation as a measure of the degree of variation in flows of different streams and by its further use has found it possible to get an approximate expression for the storage required to carry the surplus water of wet years over to dry years, which expression, in general terms, applied equally well to streams in different localities. In addition, he describes methods of estimating the probable errors in the results obtained and makes the important comment that "frank recognition of the large probable errors in many of the results cannot fail to be advantageous."[15]

The opportunities for further application of similar methods have appeared in the present writer's studies of the correlation of bacterial contents in water supplies with rainfalls upon stream watersheds and with hygienic resultants of inferior quality such as typhoid fever and diarrhoeal diseases. In these particular studies, the statistician could contribute excellent aid, since the writer is not aware of an effective method of comparing correlated phenomena in which one series of characteristics is continuous, while another is discontinuous. In addi-

* Personal communication.

tion, quantitative variations in magnitude of the values in both series are not of paramount importance, but the direction of such variations is the interesting event. The coefficient of concurrent deviations in this instance, does not appear to supply all the desiderata. An example may make our problem clearer. In the study of the daily tap water analyses of a city water supply, we find, by inspection, that the B. coli contents rise after rains on the watershed of the stream supplying the town. It is also found that such rises are masked, to varying degrees, by purification processes and by the efficiency of operation of such processes. If changes in method and efficiency of purification are brought about and the qualitative reflection of rainfalls in resultant B. coli density in tap waters is modified, how can we measure quantitatively the change in sensitiveness of tap water quality to rainfall from month to month? The data at hand for this purpose, reduced to simplest terms, are in each month B. coli values for each day (continuous series), which differ in density from day to day, and rainfall records (discontinuous series) which may give a zero value for all the days but three or four during the month. If, during the month of July, the B. coli per 100 c.c. rose from 2 to 2,000 from July 7 to July 8, following a rain of 0.8 inch on the stream on July 7, and during August the B. coli per 100 c.c. showed no jumps above 5 in spite of a number of days of rainfall of about 0.8 inch, what should be the statistical relation between the months of July and August for these particular considerations?

This paper should not be concluded without some reference to the part that the study of purification processes has played in modifying and determining the quality of water supplies and the importance therein of the mathematician's tools. It is frequently the sanitarian's problem to include in his valuation of a water's safety some definite estimate, among other things, of the efficiency of operating features involved in the treatment of such a supply. This problem has given rise to various measures of treatment efficiencies, which only recently have been subjected to rigid statistical study. As an illustration of this type of measure the percentage removal of bacteria from untreated to treated waters has persisted. Statistical objections to this measure are well known to the reader and substitutes for this measure of performance, and indirectly of quality, have been much sought after. It was long recognized that the real measure of performance should include data regarding the distribution of the efficiencies over long periods and recommendations suggesting the classification of bacterial results according to frequency distributions have done much to clarify the interpretation of treatment figures.

Further development of the same problem of plant performance

along statistical lines has been made by Wolman[23], in the study of the nature of bacterial removal in filtration plants. In this discussion, it was suggested that "the normal performance of a water filtration plant may be represented by a curve having the equation: $y=x^c$, where y and x are respectively the raw water and final effluents counts, and c is a constant for the particular plant under discussion." In other words, the tentative hypothesis was brought forth that the final effluent count, on the average, is an exponential function of the raw water count. The evaluation of "c" replaces also the unsatisfactory percentage efficiency as a more adequate measure, by using the ratio of the logarithms of the counts instead of the ratio of the actual bacterial values.

It is apparent that a measure of performance to be effective for adaptation to quality interpretation should include more than an array of its daily values, since it is the *consistency* of bacterial removal which predetermines the position of a form of treatment in the scale of the safety of a supply. Heretofore, no single unit of measure of this degree of consistency of removal has been available, although the fitting of normal performance data to the logarithmic curve of filtration supplied at least a graphic method of testing consistency.[23] If bacterial data are arranged and plotted on the probability paper already referred to in the discussion, it becomes extremely easy to obtain the values of the semi-interquartile ranges of the figures in successive steps of purification. The ratio of such values of the ranges for any two steps appears to the writer to present some promise of a real measure of the "levelling" effect of purification processes, since it measures the change produced in the frequency distribution of bacteria in passing through the treatment. The demonstration of its value may be more apparent to the reader by reference to material given elsewhere.[24]

REFERENCES

[1] Yule, G. U., An Introduction to the Theory of Statistics.

[2] Levine, Max, A Statistical Classification of the Colon-Cloacae Group, *Journal of Bacteriology*, Vol. 3, No. 3, May, 1918.

[3] McCrady, M. H., The Numerical Interpretation of Fermentation-Tube Results, *Journal of Infectious Diseases*, Vol. 17, No. 1, July, 1915.

[4] Wolman, Abel and Weaver, H. L., A Modification of the McCrady Method of the Numerical Interpretation of Fermentation-Tube Results, *Journal of Infectious Diseases*, Vol. 21, No. 3, September, 1917.

[5] McCrady, M. H., Tables for Rapid Interpretation of Fermentation-Tube Results, *The Public Health Journal* (Canada), Vol. 9, No. 5, May, 1918.

[6] Stein, Milton F., Making the B. Coli Test Tell More, *Engineering News-Record*, Vol. 78, No. 8, May 24, 1917.

[7] Stein, Milton F., On Numerical Interpretation of Bacteriological Tests, *Engineering News-Record*, Vol. 82, No. 23, June 5, 1919.

[8] Stein, Milton F., The Interpretation of B. Coli Test Results on a Numerical and Comparative Basis, *Journal of Bacteriology*, Vol. 4, No. 3, May, 1919.

[9] Greenwood, J. Junr. and Yule, G. Udny, On The Statistical Interpretation of Some Bacteriological Methods Employed in Water Analysis, *Journal of Hygiene*, Vol. 16, No. 1, July, 1917.

[10] Wells, Wm. F., The Geometrical Mean as a B. Coli Index, *Science*, N. S., Vol. 47, No. 1202, January 11, 1918.

[11] Wells, Wm. F., The Bacteriological Dilution Scale and the Dilution as a Bacteriological Unit, *American Journal of Public Health*, Vol. 9, No. 9, September, 1919.

[12] Wells, Wm. F., On a Standard System of Bacteriological Dilutions, *American Journal of Public Health*, Vol. 9, No. 12, December, 1919.

[13] Whipple, G. C., The Element of Chance in Sanitation, *Journal of the Franklin Institute*, Vol. 182, No. 1, No. 2, July and August, 1916.

[14] Wolman, Abel, Index Numbers and Scoring of Water Supplies, *Journal of the American Water Works Association*, Vol. 6, No. 3, September, 1919.

[15] Hazen, Allen, Storage to be Provided in Impounding Reservoirs for Municipal Water Supply, *Trans. American Society of Civil Engineers*, Vol. 77, p. 1539.

[16] Stein, Milton F., A Critical Study of the Bacterial Count in Water and Sewage, *American Journal of Public Health*, Vol. 8, No. 11, November, 1918.

[17] McEwen, George F. and Michael, Ellis L., The Functional Relation of One Variable to Each of a Number of Correlated Variables Determined by a Method of Successive Approximation to Group Averages: A Contribution to Statistical Methods, *Proc. American Academy Arts and Sciences*, Vol. 55, No. 2, December, 1919.

[18] Michael, Ellis L., Concerning Application of the Probable Error in Cases of Extremely Asymmetrical Frequency Curves, *Science*, N. S., Vol. 51, No. 1308, January 23, 1920.

[19] Johnstone, James, The Probable Error of a Bacteriological Analysis, *Rept. Lanc. Sea-Fish. Lab.*, 1919, No. 27. (Not read.)

[20] Hinman, J. J. Jr., American Water Works Laboratories, *Journal of the American Water Works Association*, Vol. 5, No. 2, June, 1918.

[21] Norton, J. F., Comparison of Methods for the Examination of Water at Filtration Plants, *Journal of Infectious Diseases*, Vol. 23, 1918, Pp. 344–50.

[22] Morse, Robert B. and Wolman, Abel, The Practicability of Adopting Standards of Quality for Water Supplies, *Journal of the American Water Works Association*, Vol. 5, No. 3, September, 1918.

[23] Wolman, Abel, A Preliminary Analysis of the Degree and Nature of Bacterial Removal in Filtration Plants, *Journal of The American Water Works Association*, Vol. 5, No. 3, September, 1918.

[24] Wolman, Abel and Powell, S. T., Sanitary Effect of Water Storage in Open Reservoirs, *Engineering News-Record*, Vol. 83, No. 18, October 30–November 6, 1919.